DOCUMENT RESUME

ED 109 223 TM 004 684

AUTHOR Stetz, Frank P.
TITLE A Comparison of Criterion and Norm-Referenced Assessment for the Purposes of Decision-Making.
PUB DATE 1 May 75
NOTE 13p.; Paper presented at the Annual Meeting of the New England Educational Research Organization (Provincetown, Mass., May 1, 1975); Not available in hard copy due to marginal legibility of original document

EDRS PRICE MF-$0.76 PLUS POSTAGE. HC Not Available from EDRS.
DESCRIPTORS Academic Achievement; *Comparative Analysis; *Criterion Referenced Tests; *Decision Making; Diagnostic Tests; *Norm Referenced Tests; Prognostic Tests; Program Evaluation; Test Construction; Testing; *Test Interpretation; Test Results

ABSTRACT

While much has been written on the topic of criterion-referenced testing and consequently its comparison with norm-referenced testing, measurement specialists have not as readily approached the subject of the implications involved in reporting such test information. The purposes of this paper are to (1) draw distinctions between criterion and norm-referenced assessment; (2) delineate the purposes for which uses of test information are employed; and (3) evaluate the usefulness of criterion and norm-referenced measurement in providing the necessary data for each test information use. The six major uses of test information included in the study are: prognosis, diagnosis of learning difficulty, student growth, student achievement, program evaluation and research. The six uses of test information outlined above constitute the basic requirement needs of measurement specialists. In an earlier work, Cronbach (1949), classified testing under three main headings: prognosis, diagnosis, and research; three additional uses of test information have been added: growth, achievement and program evaluation. Such an evaluation should hopefully provide school personnel with the understanding of the distinction between the types of information available from both criterion and norm-referenced testing. Such knowledge should help to promote the understanding that certain measurement information is better assessed by one type of assessment tool rather than the other depending upon the decision-making purpose in question. (Author/DEP)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED109223

TM 004 684 BEST COPY AVAILABLE

A COMPARISON OF CRITERION AND NORM-REFERENCED ASSESSMENT FOR THE PURPOSES OF DECISION-MAKING[1,2]

Frank P. Stetz
Harcourt Brace Jovanovich, Inc.

Measurement specialists have been exploring the advantages and disadvantages of criterion-referenced testing since its introduction to the educational community. As a consequence of its popularity, there have also been numerous comparisons drawn on many dimensions with norm-referenced testing. Some experts, for example, Block (1971), Glaser & Nitko (1970) and Popham & Husek (1969), contend that it is not possible to distinguish between a criterion-referenced and norm-referenced test by just looking at it. Others, Ebel (1971), Davis (1970) and Simon (1969), assert that the basic difference between the two types of instruments lies in the interpretations given to the scores of tests.

The measurement specialists do not appear to have approached the critical analysis of the implications involved in reporting such criterion-referenced test information. The position taken in this paper is that there are basic inherent differences in the intent and construction of both types of tests. Given these basic differences, it is further believed that the issue of criterion-referenced versus norm-referenced testing lies with the uses for which testing information will be employed; that is, rather than viewing and comparing such tests in an "either/or" context, one should base the evaluation upon the decision-making purpose for which testing information is needed.

---

[1] Paper presented at the annual meeting of the New England Educational Research Organization, Provincetown, Massachusetts, May 1975.

[2] The author acknowledges the helpful comments provided by Janice Scheuneman on an earlier draft of this manuscript.

Keeping the above in mind, the purposes of this paper are threefold: 1) To draw distinctions between criterion and norm-referenced assessment; 2) To delineate the purposes for which most uses of test information are employed; and 3) To evaluate the usefulness of criterion and norm-referenced measurement in providing the necessary data for each test information use. The six major uses of test information included in this study are: prognosis, diagnosis, research, program evaluation, achievement, and growth.

Characteristics of Criterion and Norm-Referenced Measurements

Criterion-referenced assessment contains certain basic characteristics commonly ascribed to such instruments: statements of specific instructional objectives are listed for particular content areas; specific mastery levels are predetermined for each objective or test; learners' responses are measured against these predetermined criterion levels. In addition, a designation is made as to which objectives have been mastered by each individual. Some optional characteristics found in many criterion-referenced assessment strategies include refinement of outcome status, summary data on the group measured and possibly the grouping of individuals for remedial instruction.

Refinement of outcome status offers more detailed information on objectives than the dichotomy, "Pass/Fail". Such categories as "Exceeded Mastery Level", "Achieved Mastery Level" and "Below Mastery Level" with entries establishing by how much scores deviate from the mastery level provide more information especially for advanced instruction or remediation purposes.

Summary data for group administrations provide useful information on

particular objectives. For example, if it is known that only 25% of the learners in a classroom mastered a particular objective, one would assume that either instruction on this objective was not effective or that the test items did not adequately measure the objective. Such summaries are gathered across individuals on particular objectives; summative information for individual examinees across all objectives on a test should not be reported. To provide a total score on diverse skills would not supply the relevant information of whether individuals have mastered specific objectives.

Summary information may also be used for the purposes of grouping for remedial instruction. If it is discovered that for a particular group of objectives certain individuals consistently do not achieve mastery, then this information may be used to set up small informal groups to provide enrichment opportunities. Thus, it is hoped that upon second testing, mastery of the objectives will be achieved.

In norm-referenced assessment explicit instructional objectives are not specified. Therefore, it is not generally possible to determine whether or not a learner has mastered particular skills. Although standardized norm-referenced tests usually report subtest scores, these subtest scores cover broad numbers of objectives under one subheading such as "Mathematics Concepts." One cannot tell which specific objectives constitute a "Mathematics Concepts" subtest without the test blueprints.

Another characteristic of norm-referenced assessment is the use of methods of item analysis to select items which discriminate best among

examinees. Items with very high or very low item difficulty are usually removed from such tests, especially those standardized on national samples of students. In a criterion-referenced testing context, the issue of discrimination is between pre and posttest results, i.e. barring prior knowledge, examinees should not be able to answer test questions before instruction but upon completion of instruction (if it has been effective) they should achieve mastery of all objectives.

A third primary characteristic of norm-referenced assessment is the comparison of an individual's score to that of others. In standardized norm-referenced tests raw scores are converted to some form of interpreted score for comparison purposes; in teacher-made tests the comparisons may be as simple as comparing Student A's total raw score with that of Student B. In any event, the emphasis is usually upon comparing total test scores and consequently assessing which examinee possesses more knowledge as identified by correct responses to test questions.

Purposes for Which Test Information Are Employed

In an earlier work, Cronbach (1949) classified testing under three main headings: prognosis, diagnosis, and research; three additional uses of test information have been added: growth, achievement, and program evaluation. While it may be argued by some that the three additional areas could justifiably be included within the framework of Cronbach's original three, discussion will hopefully show that they provide significant differences in intent to justify their individual categorization.

Prognostic testing is defined as the use of assessment techniques to predict the success of an individual in some future undertaking. An instrument is administered which has been found to successfully distinguish

those individuals who show the presence of an ability, trait, etc. which has been known to correlate highly with success on some future task. A widely known instrument of this kind is the Scholastic Aptitude Test taken by many juniors and seniors wishing to gain admission to college. The basis upon which interpretations of results are formulated rests with the admission policies set by most schools; selection of candidates is usually based upon certain criteria, for example, a score of 600 or more on both the quantitative and verbal subtests of the SAT. Selection of those candidates who show the most potential (based upon their total test scores) is the decision-making purpose in this instance.

The diagnosis of learning difficulty constitutes a second major purpose for which information is required. Information is needed to assess the level of expertise of an individual on particular skills considered important. Testing inquires into the particular patterns of correct and incorrect responses to questions, thereby providing information upon which remedial instruction can be based. The purpose of testing and the uses to which the test information will be employed, are based upon what a particular individual can and cannot do. A pre-reading test of psycho-motor ability would be an example of a diagnostic assessment tool. The teacher is interested in assessing what particular psychomotor skills a learner has acquired and on which skills more time must be spent.

Research and programs evaluation are chosen to be discussed jointly because of their apparent similarities. Testing for research purposes involves using instruments to gather quantitative data concerning achievement, aptitude, etc. for the purpose of hypothesis testing. Testing

involves drawing a representative sample of subjects who are administered tests relevant to the hypotheses in question. The researcher is usually interested in differential performance of groups or individuals on the test. A hypothetical example of testing for research purpose would be the case of an experimenter who hypothesizes that understanding mechanical relationships is correlated with tough-mindedness, practicality and careful planning. To test this hypothesis, the experimenter would draw a random sample of subjects and administer tests which appear to validly and reliably assess the variables in question, for example, such tests as the Bennett Mechanical Comprehension Test and the Thorndike Dimensions of Temperament test. The results of the analysis would be used to test the hypothesis of the relationship of mechanical comprehension and pragmatism to a particular significance level.

The use of test information for evaluation purposes is seen as a special case of testing for research. Program evaluation usually depends upon the classroom unit as the basis for sampling. Random sampling is exceedingly rare due to the cost, in both time and money, of such operations. An additional difference between testing for research and program evaluation is the purpose for which both are intended. While testing in a research context will have as its final outcome the acceptance or rejection of particular hypotheses, program evaluation results will be used for decision-making purposes concerning the program under study: whether it will be continued or suspended, whether or not increased funds will be appropriated for its further development, etc. Generally, such decision-making functions are more applied than the testing of most research hypotheses.

Program evaluation also depends upon a unit (or units) of instruction being administered and tests given to assess whether the instruction has been effective. A very common approach is to administer alternate forms as pre and posttests to insure that achievement is a product of the instruction. Attitudinal instruments may also be administered to check upon the effect that the program has on the thoughts, feelings, behaviors, etc. of the participants. For program evaluation to be most effective and provide the most meaningful data, the instruments used must be geared directly toward the instruction; i.e., the test items used must relate to the specific curriculum objectives.

Testing to assess achievement is probably the most well known purpose for which tests are utilized. An area of interest is delineated, a sample of tasks is drawn and assessment is made at a point in time of how well an individual does on those tasks. Brown (1970) lists three properties that must be present in order to classify a test as an achievement instrument:

1. The skill and content domains covered by the test can be specified and defined in behavioral terms;
2. The test does, in fact, measure these important behaviors rather than irrelevant considerations; and
3. The test takers have had equal exposure, or equal opportunities for exposure, to the material being tested (pp. 253-254).

A simplified example of an achievement measure would be a teacher-made test to assess the performance of learners on a particular unit (or units) of instruction. The teacher delineates objectives important in the instruction and attempts to construct items which directly measure those objectives of instruction.

The last purpose of testing to be explored here is the assessment of growth. Angoff (1971) defines growth as "... an increment in score associated with the passage of time ..." While many methodological considerations must be taken in account by those proposing such studies, (for example: parallel forms of a test, test score equating, distinction between growth and practice effect, reliability adequacy, etc.) such procedures if carefully carried out, offer a powerful tool in assessing the status of an individual or group over a period of time.

Examples of measurement to assess growth must necessarily deal with the methodological considerations noted above and would apply to any test used for that purpose. A test administrator must take care that the tests given at the beginning and end of the period of time in question measure the same function. (It is assumed that some intervening occurrance has taken place in the period of time between testings.) The two tests must have the same units expressed (test scores must be equated). In addition one must pay particular attention to practice effects. A partial solution would be to provide alternative forms for first and second testing insuring that a reasonable length of time has elapsed between testings.

Information Needed as the Basis for Decision-Making

The discussion of the six test information uses points out the fact that different testing situations require different kinds of information. Generally we may dichotomize test information needs into "objective-specific" and "total-comparative". Objective-specific test information appears to be assessed much better by utilizing rules for constructing criterion-referenced tests while total-comparative information is the

domain of norm-referenced testing. Although it could be possible to supply criterion-referenced test information for a norm-referenced use (and vice versa), it appears that they both provide different kinds of information more relevant to some data needs than others.

More specifically, for prognosis testing, information is needed to differentiate or sort individuals into two classifications: acceptees and rejectees. A cut-off point may be established for each particular situation; the intent is to select those individuals whose total test score meets or exceeds a certain score regardless of the individual items or patterns of items correct. A norm-referenced approach seems most suited for this purpose.

Testing for diagnostic purposes has as its main goal the discovery of patterns of correct and incorrect responses of items related to specific skills. An examination of a learner's test results should show in which areas the learner excels or is deficient. In order to answer such questions, the test items must be directly related to the skills considered important. A criterion-referenced assessment approach with test items directly relating to specific skills or objectives would appear to be the best choice in this particular situation.

In general, data gathering for the purpose of hypothesis testing relies upon the assumption that different groups or individuals possess traits, abilities, attitudes, etc. in varying amounts which when related with other variables will produce significant differences for hypotheses posed. The researcher's interest is in distinguishing among groups or individuals on their test scores thus helping to clarify their results for hypothesis testing. For such purposes of separating or categorizing individuals on the basis of test scores,

norm or comparison-referenced testing appears to be most suited. Little interest is paid to the individual responses individuals made; the emphasis is on total test score results.

To assess program effectiveness, an evaluator is most interested in determining whether the program under study was effective in reaching its program objectives. To accurately assess such a situation, the instruments used in the evaluation must be directly related to the goals of the program. Criterion-referenced testing with its commitment to items directly related to specific objectives seems well suited for this purpose.

Likewise with achievement testing, where "... skill and content domains covered by the test (should) be specified and defined in behavioral terms" and "the test (should) ... measure important behaviors rather than irrelevant considerations", criterion-referenced testing appears better suited for assessing the true state of affairs. It should be noted though that in testing to assess achievement, test administrators are often interested in drawing comparisons among classroom groups or individuals for purposes of decision making. In such instances, in addition to the primary criterion-referenced interpretations the data may also be used to extract comparisons.

Assessing growth in an individual is closely aligned to measuring achievement. The primary difference is in the time frame used for both procedures. While achievement testing is sampling behavior at a certain point of time, growth measurement attempt to sample achievement over a period of time. Given the requirements listed above that must be taken into consideration when growth studies are proposed, the tests used to assess growth should have the characteristics listed for achievement

testing; i.e., they should primarily be criterion-referenced . As was true with testing for achievement purposes, comparisons may be requested between classrooms or individuals, but the primary purposes of such measurement should not be foresaken.

In summary, the work reported herein should hopefully provide school personnel with the understanding of the distinctions between the types of information available from both criterion and norm-referenced testing. Such knowledge should help to promote the understanding that certain measurement information is better assessed by one type of assessment tool rather than the other depending upon the decision-making purpose in question.

References

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Block, J. H. Criterion-referenced measurement: Potential. School Review, 1971, 69, 289-298.

Brown, R. G. Principles of educational and psychological testing. Hinsdale, Ill.: The Dryden Press, 1970.

Cronbach, L. Essentials of psychological testing. New York: Harper & Row, 1949.

Davis, R. B. Criterion-referenced tests. In Proceedings of the Thirty-Fifth Annual Conference of Educational Records Bureau. Greenwich, Conn.: Educational Records Bureau, 1970.

Ebel, R. L. Criterion-referenced measurement: Limitations. School Review, 1971, 69, 282-288.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1970.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Simon, G. B. Comments on "Implications of Criterion-Referenced Measurement". Journal of Educational Measurement, 1969, 6, 259-260.